International Journal of Grid and High Performance Computing
Volume 12 • Issue 3 • July-September 2020

sl-LSTM:
A Bi-Directional LSTM With Stochastic Gradient Descent Optimization for Sequence Labeling Tasks in Big Data

Nancy Victor, Vellore Institute of Technology, India
https://orcid.org/0000-0002-0640-5768

Daphne Lopez, Vellore Institute of Technology, India
https://orcid.org/0000-0003-1452-2144

ABSTRACT

The volume of data in diverse data formats from various data sources has led the way for a new drift in the digital world, Big Data. This article proposes sl-LSTM (sequence labelling LSTM), a neural network architecture that combines the effectiveness of typical LSTM models to perform sequence labeling tasks. This is a bi-directional LSTM which uses stochastic gradient descent optimization and combines two features of the existing LSTM variants: coupled input-forget gates for reducing the computational complexity and peephole connections that allow all gates to inspect the current cell state. The model is tested on different datasets and the results show that the integration of various neural network models can further improve the efficiency of approach for identifying sensitive information in Big data.

KEYWORDS

Bi-Directional LSTM, Big Data, Named Entity Recognition, Sequence Labeling, sl-LSTM

INTRODUCTION

Data that primarily focuses on 3 V's: Velocity, Variety and Volume can be termed as Big Data. Some of the key sources of Big Data comprise social networking sites such as Twitter and Facebook, health-care data from various hospitals, sensor data and search logs. Big Data Analytics refers to the approach of analyzing big data sets to reveal the information which is concealed in these sets. The major benefits of Big Data Analytics include performing risk analysis, creating new revenue streams, offering tailored health care etc. But, one of the primary concerns with Big Data is preserving the privacy of data that is published (Victor, Lopez, & Abawajy, 2016).

Preserving the privacy of social network data is not an easy task as the data will usually be in unstructured formats. Most of the social network users are not aware of the privacy risks hidden behind their Facebook posts or tweets (Vallor, 2016). There can be a situation where you may not be aware that you are continuously being followed by someone who can possibly harm you by knowing about your day-to-day activities. A system that publishes social network data for public use must confirm to the principle that no details should be obtained about a particular individual from the published information even by linking the same with some external data. This is possible only by

DOI: 10.4018/IJGHPC.2020070101

applying some anonymization techniques before sharing the data for public use. In order to identify the sensitive information from unstructured data like tweets, natural language processing techniques (NLP) can be used (Victor, & Lopez, 2018).

Natural Language processing techniques enable computers to analyze human/natural language and derive meaningful information (Ma, 2006). This way of human-computer interaction proves to be efficient in the areas of automatic text summarization, sentiment analysis, named entity recognition, optical character recognition and so on. Named entity recognition (NER) refers to the technique of identifying and classifying the entities into some set of categories such as name, location etc. (Jin, Ho, & Srihari, 2009).

Sensitive information can be extracted from social network data by applying NER techniques. Recurrent neural networks can be used for NER because of the fact that it proves to be more efficient in terms of processing sequential data, as each neuron can utilize its internal memory to retain information regarding the prior input. (Camron, 2016).

RELATED WORK

Named Entity Recognition

Named Entity Recognition refers to an interesting information extraction technique in the area of machine learning, with the help of which certain types of entities can be identified using annotations. This plays a major role in giving solutions to real world queries such as whether a tweet mentions a particular person's name or location, to find the sentiments about a particular product etc. Figure 1 shows the various Named Entity Recognition and Information Extraction (IE) techniques (Christopher, Prabhakar, & Hinrich, 2008).

The efficiency of the NER approach can be evaluated using the following measures (Powers, 2011). TP, FP and FN refer to "True positives", "False positives" and "False negatives," respectively.

Precision (P): This refers to the ratio of correctly predicted entities to all the entity predictions.

$$P = \frac{TP}{TP + FP} \tag{1}$$

Recall(R): This refers to the ratio of correctly predicted entities to all the real entities.

Figure 1. NER techniques